# 研究ノート

# 相関分析

沖　津　　直

CORRELATON ANALYSIS

OKITSU Tadashi



## １．はじめに

相関は、関心のある複数の変数の間に何か関連があるかどうかを知りたいときに用いる。回帰分析と同様に、相関分析も記述的にまたは推測統計的に考察できる。統計学の発展の初期の段階では、主に生物学の分野で盛んに使われたが、現在では自然科学はもちろんだが、社会科学の経済学、経営学などでも盛んに使われている。2変数の場合たとえば、身

長と体重、ある科目の前期試験と後期試験の成績、一人当たりの食肉消費量と一人当たりの小麦消費量、所得と消費、広告費と売上高などの場合の相関をどのように考えていけばよいのか。後の2つの例は、一方が原因で他方が結果というふうに考えられるので、回帰分析でよく取り上げられる。まず、相関係数の求め方、計算の仕方を説明し、相関係数の意味について考えていきます。次に、相関係数の検定と推定を扱う。最後に、分布によらない順位相関を扱う。

## 2．相関係数の求め方

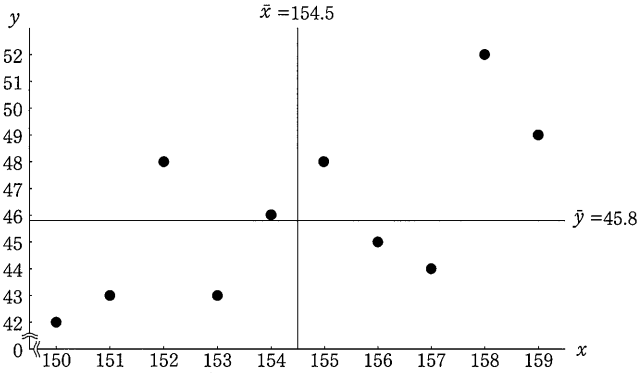
相関係数は2つの側面をもっている。ひとつは、2つの変数 $x$、$y$ 間の共変性の度合の尺度としての側面と、もうひとつは観測値の分布に対する回帰線の適合の度合の尺度としての側面である。後者の側面については、回帰分析で扱われるので、ここでは前者について論じていきたい。

2つの変数 $x$、$y$ との間に直線的関係（線形相関）があるとみとめられるとき、この関係の強さを表す尺度があると便利である。1表のような標本の大きさ10の観測値あるいは測定値が得られているとしよう。

1表　10人の男子学生の身長と体重の測定値

| NO | 身長 | 体重 | NO | 身長 | 体重 |
|---|---|---|---|---|---|
| 1 | 155 | 46 | 6 | 151 | 43 |
| 2 | 157 | 44 | 7 | 156 | 45 |
| 3 | 159 | 49 | 8 | 154 | 46 |
| 4 | 158 | 52 | 9 | 153 | 43 |
| 5 | 150 | 42 | 10 | 152 | 48 |

1図はこのデータの散布図または点相関図である。身長と体重の関係は、身長が高いほど体重も重くなるという傾向にある。また、逆に体重が重いほど身長も高いともいえ、身長と体重の関係は、単純な相互依存




1図　散布図

の関係にある。相関係数を $r$ とすると、2変数の相関係数は、次の（1）式あるいは（2）式によって求めることができる。この2つの式は同じ内容の式であり、（1）式を変形して（2）式を導くことができる。（1）式を定義式、（2）式を計算式とよばれることもある。Σは総和記号で $i$=1から $n$ を足すことであるが、ここでは省略している。

$$r=\frac{\frac{1}{n-1}\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{1}{n-1}\sum(x-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum(y-\bar{y})^2}}=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} \qquad (1)$$

$$r=\frac{n\sum xy-(\sum x)(\sum y)}{\sqrt{\left[n\sum x^2-(\sum x)^2\right]\left[n\sum y^2-(\sum y)^2\right]}} \qquad (2)$$

一般に、この $r$ は標本の大きさ $n$ の標本データから計算されたものであるから、標本相関係数と呼ばれている。この式は、イギリスの生物学者ピアソンによって創案され、単純相関係数といわれている。（1）式の分母の $\sqrt{\frac{1}{n-1}\sum(x-\bar{x})^2}$ は $x$ の標準偏差であり、$\sqrt{\frac{1}{n-1}\sum(y-\bar{y})^2}$ は $y$ の標準偏差である。分母は $x$ の標準偏差掛ける $y$ の標準偏差となっている。一方、分子の $\frac{1}{n-1}\sum(x-\bar{x})(y-\bar{y})$ は $x$ と $y$ の共分散という。したがって、相関係数は $\frac{\text{共分散}}{x\text{の標準偏差}\times y\text{の標準偏差}}$ とも書くことができる。

2表 （1）式を使って相関係数 $r$ を求める計算

| $x$ | $y$ | $x-\bar{x}$ | $y-\bar{y}$ | $(x-\bar{x})(y-\bar{y})$ | $(x-\bar{x})^2$ | $(y-\bar{y})^2$ |
|---|---|---|---|---|---|---|
| 155 | 46 | -0.5 | 0.2 | 0.1 | 0.25 | 0.04 |
| 157 | 44 | 2.5 | -1.8 | -4.5 | 6.25 | 3.24 |
| 159 | 49 | 4.5 | 3.2 | 14.4 | 20.25 | 10.24 |
| 158 | 52 | 3.5 | 6.2 | 21.7 | 12.25 | 38.44 |
| 150 | 42 | -4.5 | -3.8 | 17.1 | 20.25 | 14.44 |
| 151 | 43 | -3.5 | -2.8 | 9.8 | 12.25 | 7.84 |
| 156 | 45 | 1.5 | -0.8 | -1.2 | 2.25 | 0.64 |
| 154 | 46 | -0.5 | 0.2 | -0.1 | 0.25 | 0.04 |
| 153 | 43 | -1.5 | -2.8 | 4.2 | 2.25 | 7.84 |
| 152 | 48 | -2.5 | 2.2 | -5.5 | 6.25 | 4.84 |
| 1545 | 458 | | | 56 | 82.5 | 87.6 |
| 154.5 | 45.8 | | | | | |

（1）式を使って、早速1表のデータから $r$ を求めてみよう。1図の身長 $x$ と体重 $y$ との間には正の相関関係が認められる。しかし、直線的関係の強さの度合いは、それほど強くなく、2図の（b）ぐらいの感じである。相関係数を求める計算は、次の3表のように計算に必要な項目を拡幅して行うと便利である。この計算結果を（1）式に代入すると、

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{[\sum(x-\bar{x})^2][\sum(y-\bar{y})^2]}}=\frac{56}{\sqrt{(82.5)(87.6)}}\fallingdotseq 0.658732$$

となる。

ここで共分散を $C(x,\ y)$ で示すと、

$$C(x,\ y)=\frac{1}{n-1}\sum(x-\bar{x})(y-\bar{y})=\frac{56}{10-1}\fallingdotseq 6.222$$

となっている。また、（2）式を使っても同じ結果になることが確かめられる。この場合の計算は、3表のようになる。

以上の計算より、$n=10$、$\Sigma x=1545$、$\Sigma y=458$、$\Sigma xy=70817$、$\Sigma x^2=238785$、$\Sigma y^2=21064$ をそれぞれ（2）式に代入すると

$$r=\frac{10(70817)-(1545)(458)}{\sqrt{[10(238785)-(1545)^2][10(21064)-(458)^2]}}=\frac{560}{\sqrt{[825][876]}}\fallingdotseq 0.658732$$

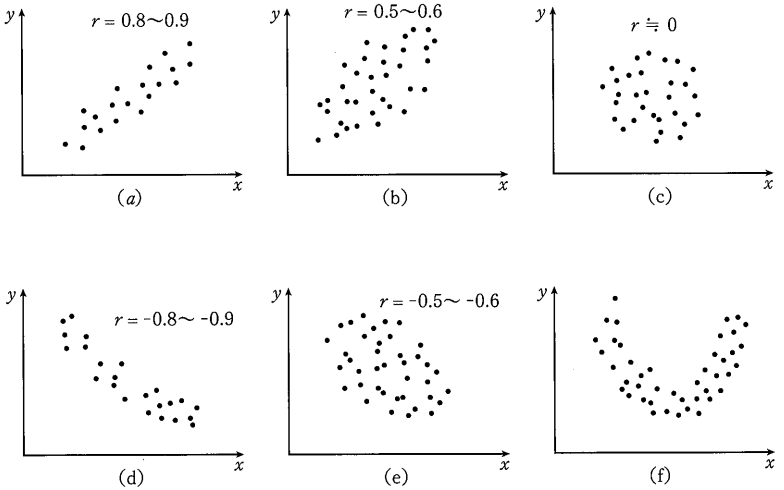
となる。このように3表のように $xy$、$x^2$、$y^2$ の項目を計算して（2）式より $r$ を求めることができる。ここで、$r\fallingdotseq 0.66$ と計算される。

3表　(2) 式を使って相関係数 $r$ を求める計算

| 身長 ($x$) | 体重 ($y$) | $xy$ | $x^2$ | $y^2$ |
|---|---|---|---|---|
| 155 | 46 | 7430 | 24025 | 2116 |
| 157 | 44 | 6908 | 24649 | 1936 |
| 159 | 49 | 7791 | 25281 | 2401 |
| 158 | 52 | 8216 | 24964 | 2704 |
| 150 | 42 | 6300 | 22500 | 1764 |
| 151 | 43 | 6493 | 22801 | 1849 |
| 156 | 45 | 7020 | 24336 | 2025 |
| 154 | 46 | 7084 | 23716 | 2116 |
| 153 | 43 | 6579 | 23409 | 1849 |
| 152 | 48 | 7296 | 23104 | 2304 |
| 1545 | 458 | 70817 | 238785 | 21064 |
| 154.5 | 45.8 | | | |

## 3．相関係数の意味

一般的に、2変数の散布図を図示してみると、2図のようになる。




2図　散布図の類型

同図の6つの類型のパターンを理解しておくことが重要である。

(a)　$x$が増すと$y$が右上方向に直線的に増加し、その直線的関係が強い。この場合、$x$と$y$との間には正の相関があり、相関の度合いは強い。

(b)　$x$が増すと$y$がやはり右上方向に増加するが、その直線的関係はそれほど強くない。この場合も、$x$と$y$との間には正の相関があるが、相関の度合（a）ほど強くない。

(c)　$x$と$y$との間には直線的関係がない。この場合、$x$と$y$との間には相関がない、または$x$と$y$とは無相関である、という。

(d)　$x$が増すと$y$は右下方向に直線的に減少し、その直線的関係が強い。この場合、$x$と$y$との間には負の相関があり、強い負の相関を示している。

(e)　$x$が増すと$y$は右下方向に直線的に減少するが、その直線的関係はそれほど強くない。この場合も、$x$と$y$との間には負の相関があるが、(d) の場合ほど強くない。

(f)　$x$があるところまで増すと$y$は減少するが、さらにそれ以上$x$が増すと、$y$は増加する。この場合、$x$と$y$との間には2次式的な関係になる。相関係数は$x$と$y$との直線関係の度合を測る尺度であるから、相関係数を計算しても無意味である。

次に、相関係数のいくつかの性質を箇条書きにしておくとつぎのようになろう。

(1)　相関係数$r$は$-1$と1の間の値をとる。すなわち、$-1 \leqq r \leqq 1$である。また、$r=+1$のとき、$x$と$y$との間には1次の関係式が成り立つ。ある定数a、bに対して、

$$y_i = \mathrm{a} + \mathrm{b}x_i \qquad \mathrm{i} = 1,\ 2,\ \cdots,\ n$$

が成立することである。$r=1$のとき、$\mathrm{b} > 0$、$r=-1$のとき、$\mathrm{b} < 0$である。

(2)　(1)で得られる$-1 \leqq r \leqq 1$の性質と前述の共分散の性質とを結びつ

けると、$r$の符号は正、負の相関を表し、$r$の値が1に近いほど正の相関が強く、$-1$に近いほど負の相関が強く、0に近ければ、$x$と$y$との間には相関がない、ということができる。

(3) $r^2\times100\%$は寄与率と呼ばれ、$y$の変動のうち$x$でどれだけの割合が説明つくかを表す数値となる。これは、回帰分析で決定係数と呼ばれている。

(4) $r$は$x$、$y$を測るときの原点の位置や尺度のとりかたに無関係である。もとのデータの値をある定数で割ったり、掛けたり、足したり、引いたりしても、$r$の値は変わらない。この性質を利用して、もとのデータを計算しやすくして$r$を求めることができる。

この例として、1表の身長データから150を引き、体重データから40を引き、$x'=x-150$、$y'=y-40$として(1)および(2)式の$x$のかわりに$x'$、$y$のかわりに$y'$とすればよい。この例の$r$を計算式から求めると、次の4表のようになる。

**4表　もとのデータからある一定数を引いて求めた相関係数の計算**

| $x'$ | $y'$ | $x'y'$ | $x'^2$ | $y'^2$ |
|---|---|---|---|---|
| 5 | 6 | 30 | 25 | 36 |
| 7 | 4 | 28 | 49 | 16 |
| 9 | 9 | 81 | 81 | 81 |
| 8 | 12 | 96 | 64 | 144 |
| 0 | 2 | 0 | 0 | 4 |
| 1 | 3 | 3 | 1 | 9 |
| 6 | 5 | 30 | 36 | 25 |
| 4 | 6 | 24 | 16 | 36 |
| 3 | 3 | 9 | 9 | 9 |
| 2 | 8 | 16 | 4 | 64 |
| 45 | 58 | 316 | 285 | 424 |

$$r=\frac{10(316)-(45)(58)}{\sqrt{[10(285)-(45)^2][10(414)-(58)^2]}}=\frac{3160-2610}{\sqrt{[(2850-2025)]\cdot[(4240-3364)]}}=\frac{560}{\sqrt{[825][876]}}\fallingdotseq0.658732$$

となって、まったく同じ答えを確認することができる。したがって、か

なりの相関性があると認められる。慣れてくれば、データの散布図を描き眺めてみることによって相関係数の度合いをある程度予想することができよう。

一般的に、データの大きさがそれほど大きくなければ（1）式あるいは（2）式でもとめることができる。しかし、測定値ないし観測値の数が非常に大きくなってくると、いままでの計算方法で処理しきれなくなってくる。そのようなとき、度数を導入した次の5表のような2次元の度数分布が必要になってくる。5表は、25人の学生の身長と体重の2変数データをまとめたものである。しかし、ここでは、手計算の都合上、度数の大きさは小さくしている。5表のような量的分類を重複させたものを、相関表（2次元の度数分布）という。相関表の行と列は、そのひとつひとつがそれぞれひとつの度数分布となっている。相関表の個々の単位はまずひとつの量的性質の大きさによっていくつかの級に分類され、その各級に分類された単位がもう一度別の量的性質の大きさによって、さらにいくつかの級に分類されている。したがって、相関表の2重の度数分布は個々の単位がもつ2つの量的性質の間の関係を示しているのであって、相関表の分析の目的は、この2つの量的性質の関係を分析することである。$x$の一定値に対する度数の和を$x$の周辺度数といい、周辺度数が示す変数$x$、$y$の度数分布を周辺度数分布という。

この表から身長が同じぐらいであっても体重にはかなりの散らばりあるいはバラツキがみられる。また、逆に同じぐらいの体重であってもその身長には散らばりがみられる。そして、散らばりがあまりない場合、$x$と$y$との結びつきは強く相関係数の値は大きくなる。データは散布図に描くことによって散らばり具合がわかるのである。

度数を導入した5表の相関表より、相関係数$r$を求めてみよう。この場合、6表の測定値の代わりに階級値に置き換えることによって、（1）式あるいは（2）式を使って求めても同じ答えが算出できる。したがって、この場合のデータを書き換えると、次のような7表のようになる。7表の計

算は、観測値を各階級の階級値に置き換えることによって、(1) 式あるいは (2) 式を使って求めても同じ答えが算出できる。さらに、ここでは原データの $x$ と $y$ からそれぞれ160, 55を差し引いたものをあらためて3列と4列に $x$ と $y$ として求めている。

$$r=\frac{25(4074)-(294)(311)}{\sqrt{[25(3908)-(294)^2][25(4537)-(311)^2]}}=\frac{101850-91434}{\sqrt{(11264)(16704)}}=\frac{10416}{\sqrt{188153856}}\fallingdotseq 0.76$$

5表　25人の学生の身長と体重のデータ

| | 160～164 | 164～168 | 168～172 | 172～176 | 176～180 | $y$の周辺度数 |
|---|---|---|---|---|---|---|
| 56～60 | 1 | 1 | | | | 2 |
| 60～64 | | 2 | 2 | 1 | | 5 |
| 64～68 | | 1 | 3 | 2 | | 6 |
| 68～72 | | | 2 | 3 | 2 | 7 |
| 72～76 | | | | 3 | 1 | 4 |
| 76～80 | | | | | 1 | 1 |
| $x$の周辺度数 | 1 | 4 | 7 | 9 | 4 | 25 |

6表　もとのデータを階級値で置き換える相関係数の求め方

| | 160～164 | 164～168 | 168～172 | 172～176 | 176～180 | 階級値 $y$ | $f$ |
|---|---|---|---|---|---|---|---|
| 56～60 | 1 | 1 | | | | 58 | 2 |
| 60～64 | | 2 | 2 | 1 | | 62 | 5 |
| 64～68 | | 1 | 3 | 2 | | 66 | 6 |
| 68～72 | | | 2 | 3 | 2 | 70 | 7 |
| 72～76 | | | | 3 | 1 | 74 | 4 |
| 76～80 | | | | | 1 | 78 | 1 |
| 階級値 $x$ | 162 | 166 | 170 | 174 | 178 | | |
| $f$ | 1 | 4 | 7 | 9 | 4 | | 25 |

このように、相関表はデータを $x$ と $y$ のそれぞれの階級値に置き換えることによって相関係数を計算できる。相関の他の例として、勉強時間と成績、肥料の量と収穫高、弁当と飲み物の販売数、そして、気温の高い夏にはビールが売れるし、気温の低い冬にはおでんが売れる、駅から遠くなればアパートやマンションの賃貸料は安くなるなども相関として考え

7表　5表のデータを2表のように並び変えて相関係数を求める計算

| $x$ | $y$ | $x$ | $y$ | $xy$ | $x^2$ | $y^2$ |
|---|---|---|---|---|---|---|
| 162 | 58 | 2 | 3 | 6 | 4 | 9 |
| 166 | 58 | 6 | 3 | 18 | 36 | 9 |
| 166 | 62 | 6 | 7 | 42 | 36 | 49 |
| 166 | 62 | 6 | 7 | 42 | 36 | 49 |
| 166 | 66 | 6 | 11 | 66 | 36 | 121 |
| 170 | 62 | 10 | 7 | 70 | 100 | 49 |
| 170 | 62 | 10 | 7 | 70 | 100 | 49 |
| 170 | 66 | 10 | 11 | 110 | 100 | 121 |
| 170 | 66 | 10 | 11 | 110 | 100 | 121 |
| 170 | 66 | 10 | 11 | 110 | 100 | 121 |
| 170 | 70 | 10 | 15 | 150 | 100 | 225 |
| 170 | 70 | 10 | 15 | 150 | 100 | 225 |
| 174 | 62 | 14 | 7 | 98 | 196 | 49 |
| 174 | 66 | 14 | 11 | 154 | 196 | 121 |
| 174 | 66 | 14 | 11 | 154 | 196 | 121 |
| 174 | 70 | 14 | 15 | 210 | 196 | 225 |
| 174 | 70 | 14 | 15 | 210 | 196 | 225 |
| 174 | 70 | 14 | 15 | 210 | 196 | 225 |
| 174 | 74 | 14 | 19 | 266 | 196 | 361 |
| 174 | 74 | 14 | 19 | 266 | 196 | 361 |
| 174 | 74 | 14 | 19 | 266 | 196 | 361 |
| 178 | 70 | 18 | 15 | 270 | 324 | 225 |
| 178 | 70 | 18 | 15 | 270 | 324 | 225 |
| 178 | 74 | 18 | 19 | 342 | 324 | 361 |
| 178 | 78 | 18 | 23 | 414 | 324 | 529 |
|  |  | 294 | 311 | 4074 | 3908 | 4537 |

ることができる。日常の仕事で、関係があるものを発見していくと、仕事の効率を高めていくことができよう。

## 4．相関係数の仮説検定および区間推定

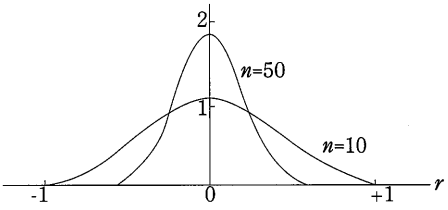
1表や5表のデータで求められる$r$の値は、繰り返し実験で同様のデータがえられるときの標本値の系列$r_1$、$r_2$、$r_3$…における最初の標本値であ

ると考えられる。これら標本値の集まりは大きさ $n=10$ あるいは $n=25$ の無作為標本を抽出して得られたものと考えられる。繰り返し実験で得られる多数の $r$ の値を度数分布表に分類し、ヒストグラムを描けば、$r$ の標本分布が得られる。相関係数の検定・推定を行うためには、$r$ の標本分布がわかっていなければならない。母集団が2変量正規分布のときには、$r$ の標本分布を知ることができる。この $r$ の標本分布は、母集団の相関係数 $\rho$ が0もしくは0でないかによって異なる。

（1） $\rho=0$ の場合

この場合、$r$ の平均と分散はそれぞれ次のようになることがわかっている。記号 $E$ は期待値を、$V$ は分散をそれぞれ表している。

$$\left.\begin{aligned} E(r)&=0 \\ V(r)&=\frac{1-r^2}{n-2} \end{aligned}\right\} \quad (3)$$




3図　$r$ の標本分布

しかも、$r$ の標本分布は、3図のように0を中心として左右対称の標本分布である。$r$ の標本分布の具体的な形は、$\rho$ と $n$ によって決まるが、ここでは $\rho=0$ であるから、同図に示すように $n$ だけに依存する。したがって、$r$ の値の確率を計算するには、$n$ だけわかればよい。付表1（出所：参考文献1p1092）に自由度 $\phi=n-2$ と有意水準に対応する $r$ の値が掲載されている。たとえば、標本の大きさが12であれば、自由度は10となる。そして、有意水準 $\alpha$ が0.05であれば、$r$ の棄却域の範囲は、

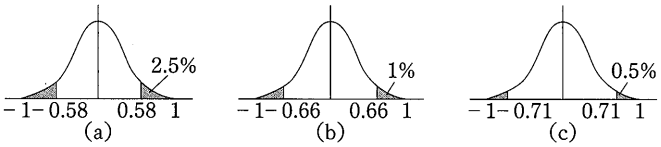
$$P(-0.5760<r<0.5760)=1-0.05$$

となる。このようにして、有意水準0.02、0.01に対応する棄却域の範囲は、

$$P(-0.6581<r<0.6581)=1-0.02$$

$$P(-0.7079<r<0.7079)=1-0.01$$

となり、これらに対応する検定の棄却域は、4図の（a）、（b）、（c）にそれぞれ示されている。




4図　$r$の標本分布

たとえば、25人の学生の身長と体重を示したデータを無作為標本として、帰無仮説$\rho=0$の検定を行ってみる。有意水準$\alpha=0.05$とする。

まず、それぞれの仮説を

$$\begin{cases} H_0：\rho=0 \\ H_0：\rho\neq 0 \end{cases}$$

と設定する。$H_0$は帰無仮説を、$H_1$は対立仮説をあらわしている。標本から求めた標本相関係数は、$r=0.76$であった。自由度$\phi=n-2=25-2=23$、$\alpha=0.05$であるから、付表1より、

$$P(-0.4227<r<0.4227)=0.95$$

であるから、$r=0.76$は3図の棄却域に存在することになります。ゆえに、帰無仮説$H_0$を棄却する。したがって、$r=0.76$と$\rho=0$との間には有意差が認められる。

以上の検定は、$t$分布表を使っても行うことができる。$r$を標準化した次の統計量

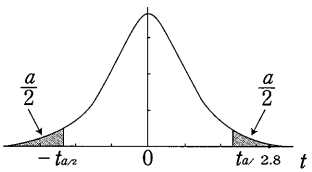
$$t=\frac{r-E(r)}{\sqrt{V(r)}}=\frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \qquad (4)$$

は、自由度$\phi=n-2$の$t$分布に従うので、$r=0.76$の$t$統計量を求めると、

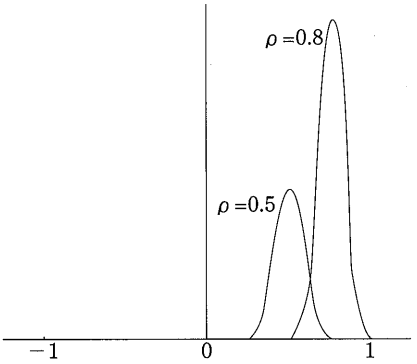
$$t=\frac{0.76}{\sqrt{\frac{1-(0.76)^2}{25-2}}}\fallingdotseq 2.77$$

これに対して、$\phi=25-2=23$、$\alpha=0.05$に対する境界値$t_{0.025}$は$t$分布表より、$t_{0.025}=1.714$であるから、

$$P(-1.714<t<1.714)=0.95$$




5図　自由度23の$t$分布




6図　$r$の標本分布

となる。5図に示すように、$t$の実現値は、棄却域に存在することになる。ゆえに、帰無仮説$H_0$を棄却することになり、前と同じ結論になる。

（2）$\rho \neq 0$の場合

$\rho \neq 0$を仮定すると、$r$の標本分布は6図に示すように対称にはならないので、付表1を使うことができない。R.A. フッシャは、$r$を次のように変換

$$Z_r = \frac{1}{2}\log_2 \frac{1+r}{1-r} = \frac{1}{2}(2.3026)\log_{10}\frac{1+r}{1-r} \qquad (5)$$

すると、この$Z$は近似的に正規分布に従うことを発見した。その平均と分散は次のとおりである。

$$\left.\begin{aligned} E(Z_r) &= Z_r = \frac{1}{2}\log_e \frac{1+\rho}{1-\rho} \\ V(Z_r) &= \sigma_r^2 = \frac{1}{n-3} \end{aligned}\right\} \qquad (6)$$

したがって、直接$r$の検定をするかわりに、間接的に$Z$の有意性検定を行う。

これまでと同様に、付表1の$n=25$の標本より求めた$r=0.76$について、$\rho=0.3$に対する検定を行ってみよう。仮説は、次のように設定する。

$$\begin{cases} H_0 : \rho = 0.3 \\ H_1 : \rho \neq 0.3 \end{cases}$$

まず、(5) 式を用いて、$r-Z$変換を行う。

$$Z_r = \frac{1}{2}(2.3026)\log_{10}\frac{1+0.76}{1-0.76} = 0.996 \fallingdotseq 1$$

$$Z_\rho = \frac{1}{2}(2.3026)\log_{10}\frac{1+0.3}{1-0.3} \fallingdotseq 0.309$$

これらの値を標準正規分布への変換公式に代入して、

$$Z = \frac{Z_r - Z_\rho}{\sigma_r} = \frac{0.996 - 0.309}{\sqrt{\frac{1}{25-3}}} \fallingdotseq 3.22$$

これに対して、標準正規分布の両すそに5％を採ったときの2つの境界点は± $Z_{0.025}=1.96$であったから、$Z$の値と正の境界点の大きさを比較すると、

$$Z=3.22 > Z_{0.025}=1.96$$

となって、棄却域に落ちていることになる。ゆえに、帰無仮説$H_0$を棄却する。明らかに、$\rho=0.3$と$r \fallingdotseq 0.76$との間には有意差があり、この標本は$\rho=0.3$の標本とは考えにくい。

最後に、$\rho$の信頼区間について述べておこう。$Z_r$が、

$$Z_r \sim N\left(Z_\rho, \frac{1}{25-3}\right) \qquad (7)$$

であるから、$\rho$の信頼区間を求めることができる。これまでの正規分布の性質から、$(1-\alpha)$の$Z_\rho$の信頼区間は、

$$P(Z_r - Z_{0.05}\sigma_r < Z_\rho < Z_r + Z_{0.05}\sigma_r) = 1 - \alpha \qquad (8)$$

より、求められる。たとえば、2表の$r \fallingdotseq 0.76$から、$\alpha=0.05$の$Z_\rho$の信頼区間を求めてみよう。

$r-Z$変換図（参考文献1のページp1093）において、$Z_\rho$の値の上側目盛と下側目盛の0.34と1.18に対応する$\rho$（上側目盛）の値を読み取ると、次のようになる。

| $Z$ | $\rho$ |
|---|---|
| 0.34 | 0.327 |
| 1.18 | 0.829 |

$$\therefore \quad 0.327 < \rho < 0.829$$

という母相関係数$\rho$の95％の信頼区間が求まる。$r-Z$変換は非常に計算が面倒であるので、既に計算したもののグラフが存在している。この

グラフを用いて、$r$の値からただちに$\rho$の信頼区間を見つけることもできる。

## 5．順位相関

（1）スピアマンの順位相関

2つの変数$x$、$y$間の依存関係は観察できるが、その分布が不明の場合、標本相関係数$r$を求めることができない。このようなとき、変数$x$、$y$間の依存関係の度合を測る順位相関が1940年に統計学者スピアマンによって開発された。これは、データの順位に基づいており、$x$と$y$の特定の分布に依存しない。変数が特定の分布に依存しない統計量をノンパラメトリックあるいは分布自由の統計量といわれている。

順位相関係数$r_s$は、次のように定義されている。

$$r_s = 1 - \frac{6\sum d^2}{n(n-1)} \qquad (9)$$

$d$は$x$と$y$の順位の差を示す。たとえば、5人の学生をランダム標本として抽出したところ、彼らの入試の順位と卒業時の順位を調べたたところ、つぎのような結果であった。

| 入試の順位 $x$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 卒業時の順位 $y$ | 2 | 4 | 1 | 3 | 5 |

8表　順位相関$r_s$を求める計算

| 入試の順位 $x$ | 卒業時の順位 $y$ | $d = x - y$ | $d^2$ |
|---|---|---|---|
| 1 | 2 | −1 | 1 |
| 2 | 4 | −2 | 4 |
| 3 | 1 | 2 | 4 |
| 4 | 3 | 1 | 1 |
| 5 | 5 | 0 | 0 |
| | | | 10 |

8表のように、右側に拡幅して計算していき、計算結果を（9）式に代

入すると、

$$r_s = 1 - \frac{6(10)}{5(5^2-1)} = 0.5$$

となる。また、もし、同順位の人が何人かいたときは、それらの順位の平均をとって、その数をそれらの人の順位とする。たとえば、上の例で、3番と4番の人の順位が同じであれば、(3+4) /2=3.5となるから、2人の順位として3.5、3.5とすればよい。あとは同様の計算を行えばよい。

ところで、$r_s$ は次のようにして導かれている。順位 $x_1$, $x_2$, …, $x_n$ は全体として、1, 2, …, $n$ に一致するから、

$$\sum x = (1 + 2 + 3 + \cdots n) = \frac{n(n+1)}{2}$$

同様に、

$$\sum y = \frac{n(n+1)}{2}$$

また、2乗和 $\Sigma x^2$ は

$$\sum x_i^2 = 1^2 + 2^2 + \cdots + n^2 = \frac{1}{6}n(n+1)(2n+1)$$

同様に

$$\sum y^2 = \frac{1}{6}n(n+1)(2n+1)$$

これらの値を普通の相関係数の公式（2）式に代入すると、スピアマンの相関係数 $r_s$ が求められる。すなわち、

$$r_s = \frac{\sum xy - \frac{1}{n}(\sum x)(\sum y)}{\sqrt{\left[\sum x^2 - \frac{1}{n}(\sum x^2)\right]\left[\sum y^2 - \frac{1}{n}(\sum y^2)\right]}} = \frac{\sum xy - \frac{1}{2}\left\{\frac{1}{2}n(n+1)\right\}^2}{\frac{1}{6}n(n+1)(2n+1) - \frac{1}{n}\left\{\frac{1}{2}n(n+1)\right\}^2}$$

上の式において、分子は、$\sum xy - \frac{1}{n}(\sum x)(\sum y) = \sum xy - n\bar{x}\bar{y}) = n\bar{x}^2 (\bar{x} = \bar{y})$

と変形できるから、これまでの値を代入すると、

$$\sum xy - n\bar{x}\bar{y} = \left[-\frac{\sum(x_i - y_i)^2}{2} + \frac{\sum x_i^2 + \sum y_i^2}{2}\right] - n\bar{x}^2$$

ここで、$d = x - y$ とすると、上の式は

$$= -\frac{\sum d_i^2}{2} + \sum x_i^2 - n\bar{x} = -\frac{\sum d^2}{2} + \frac{n(n^2-1)}{12}$$

一方、分母の $\sum x^2 - \frac{1}{n}(\sum x)^2$、$\sum y - /n(\sum y)$ は、それぞれ

$$\sum x^2 - \frac{1}{n}(\sum x)^2 = \sum x^2 - n\bar{x} = \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} = \frac{n(n+1)}{12}$$

$$\sum y^2 - \frac{1}{n}(\sum y)^2 = \sum y^2 - n\bar{y}^2 = \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)}{4} = \frac{n(n+1)}{12}$$

であるから、結局上の $r_s$ の公式は

$$r_s=-\frac{\sum d^2}{2}+\frac{n(n^2-1)n(n+1)}{12}/12=1-\frac{6\sum d^2}{n(n+1)}$$

とまとめられて、(9) 式の形になる。(9) 式の $r_s$ は、順位が同一方向に完全に一致すれば、$r_s=1$ となり、逆に反対方向に完全に一致すれば $r_s=-1$ となる。したがって、

$$-1\leqq r_s\leqq 1$$

である。

以上のように、相関分析の見方と考え方について考察してきました。相関分析は、相関係数の大きさによって役立つ度合が違ってくる。散布図のところで説明したように、相関係数が大きくなければ利点が生まれない。

この小稿ではそれほど多くのことを示すことができなかったが、まず、前半では記述統計学の立場から相関係数の求め方と意味を説明し、後半では推測統計学の立場から相関係数の仮説検定と区間推定を扱いました。相関係数を使用するとき、注意しなければならないことが2、3ある。第1は、因果関係がはっきりしない点です。明らかに因果関係がわかる場合もあります。どちらが原因で、どちらが結果なのか、はっきりしない場合もある。身長と体重の相関係数が大きくても、その結果から身長が高いから体重が重いのか、体重が重いので身長が高いのかはっきりしません。第2は、みせかけの相関あるいは擬似相関と言われているもので、既存の知識、考え方では合理的な説明のつかないまったく無意味な相関が存在するということである。計算した相関係数が高いからといって、そのまま相関関係が強いとは単純に言えない場合がある。2変数がともに別の変数によって影響を受け、そのために2変数間に数学的関係が生じているかもしれないのである。もし、相関係数が2変数間の関係について微妙な情報を示唆するようなときには、その取り扱いには慎重を期さなければならない。相関係数を有効に利用するには、適用分野の知識も精通している必要がある。

## 参考文献

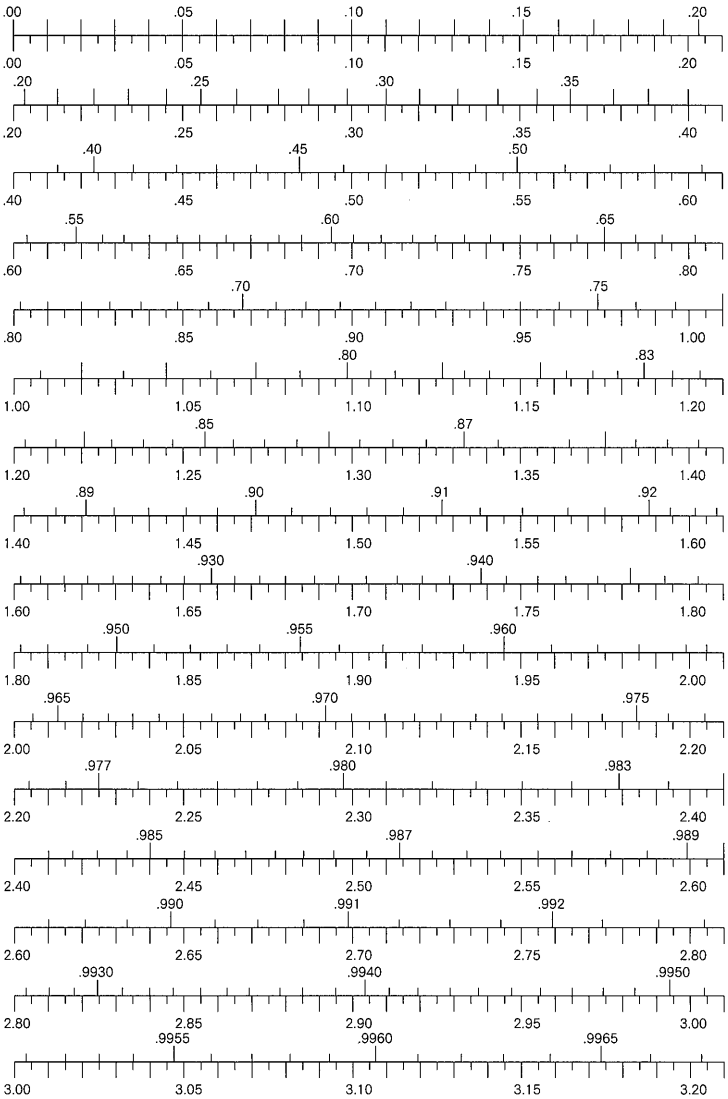
1. Statistics taro Yamane 3ed. Harper & Row 1973
2. Elementary statistics P.G.Hoel 4ed. John Wiley Sons, Inc. 1975

## 付表1　$r$（単純相関関数）の値

| $\phi$ | .1 | .05 | .02 | .01 | .001 |
|---|---|---|---|---|---|
| 1 | .98769 | .99692 | .999507 | .999877 | .9999988 |
| 2 | .90000 | .95000 | .98000 | .990000 | .99900 |
| 3 | .8054 | .8783 | .93433 | .95873 | .99116 |
| 4 | .7293 | .8114 | .8822 | .91720 | .97406 |
| 5 | .6694 | .7545 | .8329 | .8745 | .95074 |
| 6 | .6215 | .7067 | .7887 | .8343 | .92493 |
| 7 | .5822 | .6664 | .7498 | .7977 | .8982 |
| 8 | .5494 | .6319 | .7155 | .7646 | .8721 |
| 9 | .5214 | .6021 | .6851 | .7348 | .8471 |
| 10 | .4973 | .5760 | .6581 | .7079 | .8233 |
| 11 | .4762 | .5529 | .6339 | .6835 | .8010 |
| 12 | .4575 | .5324 | .6120 | .6614 | .7800 |
| 13 | .4409 | .5139 | .5923 | .6411 | .7603 |
| 14 | .4259 | .4973 | .5742 | .6226 | .7420 |
| 15 | .4124 | .4821 | .5577 | .6055 | .7246 |
| 16 | .4000 | .4683 | .5425 | .5897 | .7084 |
| 17 | .3887 | .4555 | .5285 | .5741 | .6932 |
| 18 | .3783 | .4438 | .5155 | .5614 | .6787 |
| 19 | .3687 | .4329 | .5034 | .5487 | .6652 |
| 20 | .3598 | .4227 | .4921 | .5368 | .6524 |
| 25 | .3233 | .3809 | .4451 | .4869 | .5974 |
| 30 | .2960 | .3494 | .4093 | .4487 | .5541 |
| 35 | .2746 | .3246 | .3810 | .4182 | .5189 |
| 40 | .2573 | .3044 | .3578 | .3932 | .4896 |
| 45 | .2428 | .2875 | .3384 | .3721 | .4648 |
| 50 | .2306 | .2732 | .3218 | .3541 | .4433 |
| 60 | .2108 | .2500 | .2948 | .3248 | .4078 |
| 70 | .1954 | .2319 | .2737 | .3017 | .3799 |
| 80 | .1829 | .2172 | .2565 | .2830 | .3568 |
| 90 | .1726 | .2050 | .2422 | .2673 | .3375 |
| 100 | .1638 | .1946 | .2301 | .2540 | .3211 |

## 付表2　$z$ と $r$ の対応する値

例　$r$：上側目盛　$r = 0.7$,　$z$：下側目盛　→　$z = 0.867$




（本学経営学部教授）